ISSN 1682-296X (Print)
ISSN 1682-2978 (Online)

# Bio Technology

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Biotechnology 12 (6): 225-235, 2013
ISSN 1682-296X / DOI: 10.3923/biotech.2013.225.235


# Development, Characterization and Cross-species Transferability of Expressed Sequence Tag-simple Sequence Repeat (EST-SSR) Markers Derived from Kelampayan Tree Transcriptome

[1]P.S. Lai, [1]W.S. Ho and [2]S.L. Pang
[1]Forest Genomics and Informatics Laboratory (fGiL), Department of Molecular Biology,
Faculty of Resource Science and Technology, Universiti Malaysia Sarawak,
94300, Kota Samarahan, Sarawak
[2]Applied Forest Science and Industry Development (AFSID),
Sarawak Forestry Corporation, 93250 Kuching, Sarawak

**Abstract:** *Neolamarckia cadamba* (or locally known as kelampayan) is an important fast growing plantation tree species that confers various advantages for timber industry as a strategy for reducing the logging pressure on natural forests for wood production to an acceptable level. Hence, attempts were made to develop a set of EST-SSR markers for kelampayan trees based on the EST sequences of kelampayan (NcdbEST) and further assessed the polymorphisms and transferability of the markers to other species. In this study, 155 (2.34%) out of 6,622 EST sequences which contain 232 SSRs were mined from NcdbEST. Of these, 97 ESTs were assigned with putative functions and gene ontology terms. Eighteen EST-SSR markers were developed according to the criteria, and further characterized and validated by using 50 individuals of kelampayan from two selected mother trees. The markers exhibited a considerable high level of polymorphism in kelampayan trees with an average of 4.17 and 4.11 alleles per locus, and PIC values of 0.465 and 0.537, respectively for mother trees T1 and T2. Parentage assignment analysis suggests a high probability for kelampayan trees to be predominantly outcrossed. The transferability rate was ranging from 16.7-94.4% among the five cross-genera species of kelampayan. The present study is the first report of the development of EST-SSR markers in kelampayan. These markers will be valuable genomic resources that could pave the way for exploiting the genotype data for comparative genome mapping, association genetics, population genetics studies and molecular breeding of kelampayan and other indigenous tropical tree species in future.

**Key words:** *Neolamarckia cadamba*, kelampayan, expressed sequence tag-simple sequence repeat, EST-SSR, transferability, paternity assessment

## INTRODUCTION

*Neolamarckia cadamba* or locally known as kelampayan is a large, deciduous and fast growing tree species that gives early economic returns within 8-10 years. Under normal conditions, it attains a height of about 17 m and diameter of 25 cm at breast height (dbh) within 9 years. Depending upon the soil condition, a 10-year old tree has a diameter of 50 cm which yields 2.5-3 $m^3$ wood (Soerianegara and Lemmens, 1993). It is one of the best raw materials for the plywood industry and also serves as raw material for the pulp and paper industry. Kelampayan can also be used as a shade tree for dipterocarps line planting while its leaves and bark can be applied in medicine (Joker, 2000). The dried bark can be used to relieve fever and as a tonic while the aqueous extract of the leaves was found to be used as analgesic to reduce pain and inflammation. Due to its multipurpose function and utility as well as the short rotation period, this species is now selected for planted forest development in Malaysia, especially in Sarawak (Ho *et al*., 2010). However, little is known about the molecular markers developed for kelampayan trees.

Simple sequence repeats (SSRs) are a small array of tandemly arranged bases (1-6 bp) which dispersed evenly throughout the genome, occurring in both coding and non-coding region of all eukaryotic organisms (He *et al*., 2003). SSRs are more advantageous over many other markers as there are highly abundant, polymorphic, co-dominant inheritances, analytical simple, readily

**Corresponding Author:** Ho Wei Seng, Forest Genomics and Informatics Laboratory (fGiL), Department of Molecular Biology, Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 94300, Kota Samarahan, Sarawak, Malaysia Tel: 60 82-581000 Fax: 60 82-583160

transferable and high reproducibility (Yasodha *et al*., 2008; Ho *et al*., 2006). It has been widely utilized in plant genomic studies, especially in large-scale breeding programmes. However, the development of SSR markers is expensive, labour intensive, time consuming and species specific (Liang *et al*., 2009). Hence, much of the researchers shift their attention to the development of EST-SSR markers which is more transferable across taxonomic boundaries due to the residence in the transcribed regions and the markers may lead to the identification of genes controlling economical important traits (Liewlaksaneeyanawin *et al*., 2004; Feng *et al*., 2009). Although, EST-SSR markers are less polymorphic than genomic SSR markers, they are more informative, more cost-effective, less time-consuming, having lower frequency of null alleles and higher transferability. Compared to genomic SSRs, EST-SSRs are more useful in assessment of functional diversity and comparative mapping while genomic SSRs are superior for fingerprinting and variety identification studies (Varshney *et al*., 2005).

EST-SSR markers have been developed in many crop species, herbs species, ornamental species and forest tree species, these include coffee (Poncet *et al*., 2006), rubber trees (Feng *et al*., 2009), Clementine (Luro *et al*., 2008), *Eucalyptus* (Yasodha *et al*., 2008), peanut (Liang *et al*., 2009) and sugarcane (Marconi *et al*., 2011). Research done by Yu *et al*. (2004) revealed the foundation of the markers which might be useful for the selection of protein which functioned in signaling during cereal seed germination, meanwhile Feng *et al*. (2009) had successfully characterized the variations of the important control genes, such as high-yielding, wind-resistant, fast-growing and cold-tolerance in rubber trees. Germplasm evaluation using EST-SSR markers might enhance the role of genetic markers by assaying the variation in transcribed and known-function genes and tracking for the desirable traits in large-scale breeding programmes (Varshney *et al*., 2005). EST-SSR markers are also potentially transferable among closely related species or genera due to the conservation of gene-associated sequences (Luro *et al*., 2008). Hence, this assists in understanding gene organization and expression, and evolutionary analysis of rare, endangered or invasive plant species (Liewlaksaneeyanawin *et al*., 2004; Varshney *et al*., 2005; Ellis and Burke, 2007; Pashley *et al*., 2006).

The development of EST-SSR markers assist in addressing questions related to conservation and tree improvement programmes. For instance, the markers assist in rapid selection of the desired functional genes which also enhance breeding efficiency and reduce breeding period and field workload (Feng *et al*., 2009). Overall, the development of EST-SSR markers would greatly increase in the speed of the plant breeding cycle, thereby reducing the breeding and production costs and accelerating the production of elite genotypes or clones into market. Hence, the objectives of the study were (1) *in silico* analysis on the frequency and distribution of EST-SSR markers in kelampayan tree transcriptome (NcdbEST), (2) To develop the EST-SSR markers specific for genotyping kelampayan trees from kelampayan tree transcriptome (NcdbEST), (3) To determine the characteristics and polymorphisms of each newly developed EST-SSR markers for kelampayan and (4) To evaluate the transferability of these newly developed EST-SSR markers to other species. Results obtained from this study will provide useful genomics information and resources for future breeding and improvement of kelampayan trees.

## MATERIALS AND METHODS

**Plant materials and DNA isolation:** Two mother trees were selected from a kelampayan seed production area in Ravernscourt Camp, Lawas, Sarawak. Both mother trees (T1 and T2) are separated in a distance of 10 km apart. Fresh young leaf tissues and fruits were collected from the respective mother trees in 2010 and in subsequent year, fresh young leaves were collected from 25 progenies or seedlings of each selected mother tree. These seedlings with known maternal genotype (half-sib families) were chosen in the present study in order to evaluate the parentage assignment of EST-SSR loci in kelampayan. A sterile micropipette tip [white (1-10 μL)] was used to punch the plant leaves to ensure that uniform sizes of the tissue were captured. Sterile micropipette tip was pressed on the surface of the leaf to punch the leaf. Six captured leaf discs were immediately released into the PCR reaction tubes which containing 50 μL of Extraction Buffer. The leaf discs were then incubated together with Extraction Buffer (EB) at 95°C for 10 min in PCR machine. After the incubation, the solution in the PCR tubes was mixed by inverting and tapping. 120 μL of Dilution Buffer (DB) was added to the incubated solution and stored at -20°C for further analysis.

**Data mining of EST-SSR from NcdbEST:** EST sequences were obtained from the kelampayan expressed sequence tag database (NcdbEST). A total of 6,622 ESTs with average edited length of 478 bp were generated through high-throughput 5' EST sequencing of cDNA clones derived from developing xylem tissues (Ho *et al*., 2010). These ESTs were further assembled to generate 4,728 xylogenesis unigenes with 2,100 consensus contig

sequences (average of 621 bp) and 2,628 singletons (average of 723 bp). SSRIT software (http://www.gramene.org/db/markers/ssrtool) was used to search for SSRs from the ESTs. The search criteria are: a minimum length of 12 bp, at least 3 repeat units for tetra-, penta- and hexa-nucleotide SSRs. Dinucleotide repeats such as AT/TA, CT/GA are treated as the same type of repeat motif.

***In silico* EST-SSR analysis and primer design:** Gene Ontology (GO) term annotation and functional based analysis was performed by using the Blast2Go (www.blast2go.de/), a sequence-based tool to assign GO terms and putative function by comparison with the non-redundant sequence database at NCBI, the threshold for the expect value (E-value) used is $10^{-10}$. The gene ontology numbers for the best homologous hits were used to find molecular function, cellular component, and biological process ontology for these sequences. The forward and reverse primers for non-redundant EST-SSRs were designed using the Primer Premier 5.0 software (Premiere Biosoft, USA) based on the following core criteria: (1) GC content between 40 and 60%, (2) Annealing temperature ($T_a$) between 52 and 65°C , (3) Primer length ranging from 18 bp to 24 bp, (4) Expected products size ranging from 100-400 bp. Formation of hairpin and primer dimmers were avoided. The start and the end positions of the SSRs should be at least 50 bases from the 3' and 5' ends of sequence, respectively. The sequences contain either too little DNA sequence flanking the repeat region or inappropriate for primer design were eliminated.

**EST-SSR assay and DNA sequencing:** PCR was carried out using Mastercycler Gradient PCR (eppendorf, Germany). The use of 25 µL of PCR reaction mixture which contain 1 µL of the DNA template; 1x PCR buffer; 0.2 mM each of dNTPs (dATP, dCTP, dTTP and dGTP); 2.5 mM $MgCl_2$; 10 pmol of each forward and reverse primer; 0.5 unit of *Taq* DNA polymerase (Invitrogen, Brazil) and $ddH_2O$ was prepared. The PCR amplification was carried out under the thermal cycling profile as follow: one cycle of initial denaturation at 94°C for 3 min followed by 35 cycles of 30 sec of denaturation at 94°C, 45 sec of annealing at 55°C and 1 min of extension at 72°C and, one cycle of final extension at 72°C for 5 min. The PCR products were checked using MetaPhor® agarose (Lonza, USA) stained with GelStar® Nucleic Acid Gel Stain (Lonza, USA). PCR amplicons were purified using QIAquick® Gel Extraction Kit (QIAGEN, Germany) and then ligated into pGEM®-T Easy Vector System (Promega, USA). Colony PCR with M13 forward and reverse sequence primers were performed to identify the presence of positive clones. After that, plasmids DNA from positive clones were isolated and purified using the Wizard® Plus SV Minipreps DNA Purification System (Promega, USA) according to the manufacturer's protocol. The purified plasmids were outsourced for sequencing and the sequencing was conducted by using 3730XL DNA Analyzer (Applied Biosystems, USA) and BigDye version 3.1 (Applied Biosystems, USA). Chromas Lite version 2.01 (Technelysium Pty Ltd., Australia) was used to view the raw ABI-formatted chromatogram reads and the vector sequences were removed accordingly.

**Cross-taxon analysis:** The cross genera transferability of all scorable SSR loci was evaluated using 5 species from different genera of Rubiaceae: *Ixora caesia*, *Gardenia jasminoides*, *Mussaenda erythrophylla*, *Morinda citrifolia* and *Coffea canephora*. The percentage of transferability was calculated for each species according to the fragments detected.

**Statistical analysis:** The scored data from polymorphic loci was used to calculate Polymorphism Information Content (PIC) and Neighbor-joining values in order to construct dendrogram by using PowerMarker 3.25 software (Liu and Muse, 2005). Dendrogram was graphically displayed by using MEGA version 3.1 (Kumar *et al*., 2004). The Hardy-Weinberg equilibrium (HWE) and observed heterozygosity and expected hetrozygosity were calculated by using POPGENE 1.31 software (Yeh *et al*., 1997). The presence of null alleles and the estimated null allele frequency (r) were estimated by using Micro-Checker Version 2.2.3 (Van Oosterhout *et al*., 2004).

## RESULTS AND DISCUSSION

**Frequency and distribution of EST-SSRs:** A total of 155 out of 6,622 ESTs of kelampayan were used to evaluate the presence of SSR motifs and 232 SSRs were successfully identified from these ESTs. The frequency of EST containing SSRs in the NcdbEST was 2.34%. This result was in the range of 1.5-4.7% reported by Kantety *et al*. (2002) for barley, maize, rice, sorghum and wheat. In general, the frequency of SSRs present in ESTs would be approximately 5% when the minimum length was 20 bp. The reported frequencies of EST-SSRs for various species were varied due to the microsatellite search tools and the criteria used for microsatellites identification (Varshney *et al*., 2005). From the EST-SSR sequences, 60 (38.70%) ESTs contained more than one SSR while the rest (61.29%) contained only 1 SSR. Analysis of SSR

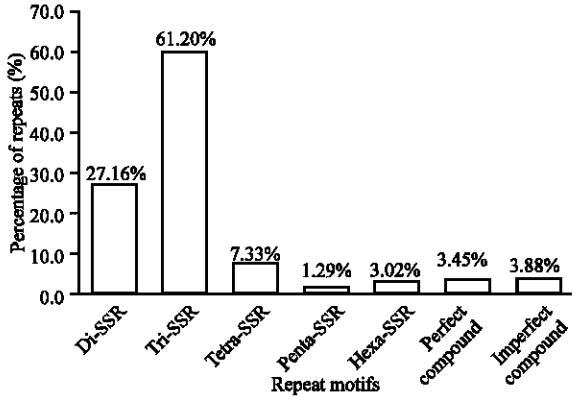


Fig. 1: Bar chart of the frequency and distribution of EST-SSRs in kelampayan tree transcriptome (NcdbEST)

Table 1: Summary of *in silico* mining of unigene sequence of kelampayan

| Characteristics | No. (%) |
|---|---|
| **Total EST sequences in database** | 6,622 |
| • EST sequences contain SSRs | 155 (2.34) |
| • ESTs contain only 1 SSR | 95 (61.29) |
| • ESTs contain more than 1 SSR | 60 (38.70) |
| **No. of SSRs found** | 232 |
| • Simple repeats | 197 |
| • Compound repeats | |
| • Perfect compound repeats | 8 |
| • Imperfect compound repeats (interval≤6 bp) | 9 |

motifs revealed that the proportion of SSR unit size was not evenly distributed. The summary of the distribution of EST-SSRs in NcdbEST is shown in Table 1.

With the exception of mononucleotide repeats, 142 SSRs (61.20%) were tri-nucleotide repeats (Tri-SSRs), 63 (27.16%) were di-nucleotide repeats (Di-SSRs), 17 (7.33%) were tetra-nucleotide repeats (Tetra-SSRs), 7 (3.02%) were hexa-nucleotide repeats (Hexa-SSRs), and 3 (1.29%) were penta-nucleotide repeats (Penta-SSRs) (Fig. 1). Furthermore, eight perfect compound repeats were detected and nine SSRs which were separated by less than six nucleotides were categorized as imperfect compound repeats. Of the total 232 SSRs found in these ESTs, 197 (84.91%) contained simple repeat motif while 35 (15.09%) represented compound motif types.

The most common dinucleotide repeats was AG/TC, accounted for 29.79%, followed by AT/TA (27.66%) and AC/TG (19.15%). In other plant species, such as wheat, rice, maize and soy bean studies, AG exhibits as the dominant EST-SSR dinucleotide repeat (Gao *et al*., 2003). CG/GC repeat was not found in NCdbEST and it is reported to be rarely found in many plant and animal genomes. This similar result was also observed in loblolly and spruce (Berube *et al*., 2007), rubber (Feng *et al*., 2009) and bean (Blair *et al*., 2011). However, GC-rich motifs are having high frequency in monocots plants and the GC-rich trinucleotide repeats are dominated in wheat, rice and maize (Gao *et al*., 2003).

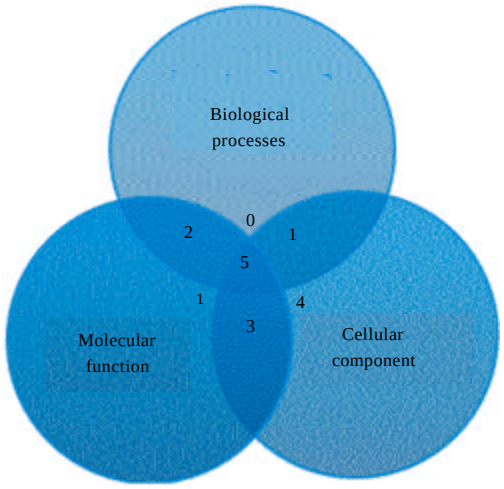


Fig. 2: Summary of Gene Ontology (GO) annotation

The motif for trinucleotide repeats was mainly CGA/GCT (15.52%), CCA/GGT (8.62%) and AAG/TTC (6.03%). CGA/GCT was not found in most of the species that had been reported while the third most abundant repeat motif, AAG/CTT was the most abundant motif for tulip tree (Xu *et al*., 2010), peanut (Liang *et al*., 2009) and rubber tree. Open Reading Frame (ORF) region are known to have higher CG percentage (Li *et al*., 2004), therefore, most of the trinucleotide repeats in kelampayan is predicted to be located in ORF region due to the high percentage of CGA/GCT. Furthermore, trinucleotide repeats are found to be relevant to certain genes with important function. CCG repeats are found to be involved in gene such as stress resistance, transcription regulation and metabolic enzyme biosynthesis meanwhile AAC repeats are related to the important agronomical traits of wheat (Gao *et al*., 2003). There were 2.16% of CCG/GGC and 1.29% of AAC/TTG repeats found in kelampayan and these sequences might be useful for gene discovery and association genetics studies in the future (Lau *et al*., 2009; Ho *et al*., 2011; Tchin *et al*., 2011, 2012).

**Functional annotation:** To explore the potential utility of the EST-SSRs, all the sequences were annotated with Gene Ontology (GO) terms, which used for gene functional annotation and classification. From 155 EST sequences which contain SSRs, 97 (62.58%) were assigned with putative functions and GO terms. A total of 37.42% of the EST-SSR sequence did not show any significant homology to the NCBI non-redundant sequence database; this result is lower than 45.82% in cacao (Lima *et al*., 2010). Of the 97 known functional ESTs, 88 were associated with gene belonging to biological process, 79 with cellular component and 75 with molecular function (Fig. 2). Many ESTs were associated

Table 2: Information of polymorphic primer pairs of kelampayan EST-SSRs

| Locus | Repeat motif | Accession No. | Ta (°C) | Expected size (bp) | Primer sequence | Marker validation: Mother tree T1 Allele No. | Mother tree T1 PIC | Mother tree T2 Allele No. | Mother tree T2 PIC |
|---|---|---|---|---|---|---|---|---|---|
| NCS01 | $(GCA)_7$ | JX446371 | 62.0 | 204 | 5' GCTCTTCTTCATCATCTCC 3'<br>5' GACCACCCCACTTCTTCC 3' | 5 | 0.579 | 5 | 0.603 |
| NCS02 | $(GGA)_6$ | JX446372 | 53.9 | 376 | 5' CTTTAGTGGGTGTAACGAGC 3'<br>5' ATCTTCCCTTCTGCTCCC 3' | 2 | 0.071 | 2 | 0.335 |
| NCS03 | $(AG)_{17}$ | JX446373 | 57.4 | 106 | 5' TGCCAGCCAGAGGAGTAG 3'<br>5' CGAGCAAAGACAGCAATG 3' | 5 | 0.670 | 4 | 0.689 |
| NCS04 | $(ATC)_7$ | JX446374 | 57.4 | 310 | 5' GCAGCTCGTCGATCTCAG 3'<br>5' ATACAGCCCAACCCAACA 3' | 3 | 0.576 | 2 | 0.316 |
| NCS05 | $(ATC)_{11}$ | JX446375 | 53.9 | 286 | 5' TCCTTCGCTTCTTCTTCG 3'<br>5' TGTGCCACTGGGATTCTG 3' | 10 | 0.857 | 10 | 0.876 |
| NCS06 | $(AGG)_6$ | JX446376 | 60.1 | 216 | 5' TGAGCAAGGCAAAGACTAAG 3'<br>5' CTGGTTCATCGCTGTCCT 3' | 8 | 0.801 | 7 | 0.762 |
| NCS08 | $(CCT)_8$ | JX446378 | 62.0 | 249 | 5' CCAACCATCTCCAACAACC 3'<br>5' CTGTGAAACTTTTGCCTCCA 3' | 4 | 0.534 | 4 | 0.641 |
| NCS09 | $(AAG)_6$ | JX446379 | 57.4 | 183 | 5' GCACTGATTGGACGACTGA 3'<br>5' GTTACTCCGTTTCGTGGG 3' | 1 | 0.000 | 2 | 0.255 |
| NCS10 | $(AGA)_9$ | JX446380 | 62.0 | 228 | 5' CGGAGTCTACTGAGGTTTCG 3'<br>5' TTCGCCTACTCTGCTTTCC 3' | 5 | 0.593 | 5 | 0.712 |
| NCS12 | $(CT)_{12}$ | JX446382 | 61.0 | 260 | 5' GACCAAAACCAACTTCCAA 3'<br>5' ATCTGACAATGGAGGACGA 3' | 6 | 0.681 | 6 | 0.737 |
| NCS13 | $(GA)_9$ | JX446383 | 55.8 | 165 | 5' GCCTGTGGTGTCATTGGT 3'<br>5' CGAATCACTACAAGGAGCAG 3' | 3 | 0.469 | 2 | 0.370 |
| NCS14 | $(TTC)_6$ | JX446384 | 55.8 | 331 | 5' CTTCTGTTCCCGTCGTCC 3'<br>5' GAGGGTCTCCATCTCATCG 3' | 4 | 0.480 | 5 | 0.746 |
| NCS15 | $(ATC)^7$ | JX446385 | 54.9 | 326 | 5' GCTATGAACAATCCGAAGG 3'<br>5' AAACTCCTGCCTGACTGC 3' | 4 | 0.620 | 2 | 0.316 |
| NCS16 | $(AGA)^6$ | JX446386 | 54.9 | 162 | 5' GTGGATGGTGAAAGCGACT 3'<br>5' CCTACTATTCATCAGCCTTTGG 3' | 1 | 0.000 | 1 | 0.000 |
| NCS17 | $(AAT)^6$ | JX446387 | 54.9 | 331 | 5' ACATAGGCTTATCAGAGGTCC 3'<br>5' CCCAAATCCACCTCACAT 3' | 1 | 0.000 | 2 | 0.350 |
| NCS18 | $(ATA)^7$ | JX446388 | 54.9 | 201 | 5' GGCTCTACAAATACATCACCTG 3'<br>5' CGGATTGACTTCTTGCTCTT 3' | 5 | 0.514 | 6 | 0.657 |
| NCS20 | $(TG)^{11}$ | JX446390 | 54.9 | 251 | 5' CAGATGAAACGGCACAGC 3'<br>5' CTCATTTCCCATTTCCACC 3' | 3 | 0.468 | 3 | 0.574 |
| NCS22 | $(CAG)^6$ | JX446392 | 55.8 | 282 | 5' GGCAGCAAATTAAAGACCA 3'<br>5' CAAAATCCCATCTGCTACTG 3' | 5 | 0.464 | 6 | 0.735 |
| | | | | | Mean | 4.17 | 0.465 | 4.11 | 0.537 |

with 2 or more category. For instance, 54 ESTs were associated with three major gene ontology categories, 13 ESTs with biological processes and cellular components, 21 ESTs with biological processes and molecular functions and 3 ESTs with cellular components and molecular functions. The number of ESTs with only one known function of a biological process, cellular component, or molecular function was 0, 4, and 1, respectively.

Primary metabolic process (16.67%) was the most dominant group annotated in the biological process category (Fig. 3a). It was followed by the cellular metabolic process at 13.33%, macromolecule metabolic process at 11.67% and biosynthetic process at 10.56%. Nucleic acid binding (33.33%) was the most dominant group in molecular function category, followed by hydrolase activity (20.37%) and protein binding (18.52%) (Fig. 3b). With regard to the cellular component, 25% were assigned to plastid followed by nucleus (22%) and mitochondrion (16%) (Fig. 3c). The EST-SSR sequences were assigned to a wide range of gene ontology categories and this indicates that these sequences are highly informative for genetic diversity and marker-assisted selection studies.

**EST-SSR markers development:** Some EST-SSR sequences were eliminated for primer design due to the following reasons: (1) Microsatellite sequences were too short, (2) Microsatellites were located near the beginning and the end (within 50 bp), or (3) The flanking sequences were not unique. Out of 24 designed primers, 18 were successfully used for PCR amplification. A total of 75 alleles were detected in mother tree T1 and 74 alleles were detected in mother tree T2, with an average of 4.17 and 4.11, respectively (Table 2). The total number of alleles ranged from 1 (NCS16) to 13 (NCS05) with an average of 4.78 alleles per locus. This value is higher than the results obtained from other plant species, such as, 3.12 in rubber (Feng *et al*., 2009), 3.92 in cacao (Lima *et al*., 2010) and 2.1 in peanut (Liang *et al*., 2009).

A total of 83.33% (15/18) and 94.44% (17/18) loci were polymorphic in mother tree T1 and T2, respectively

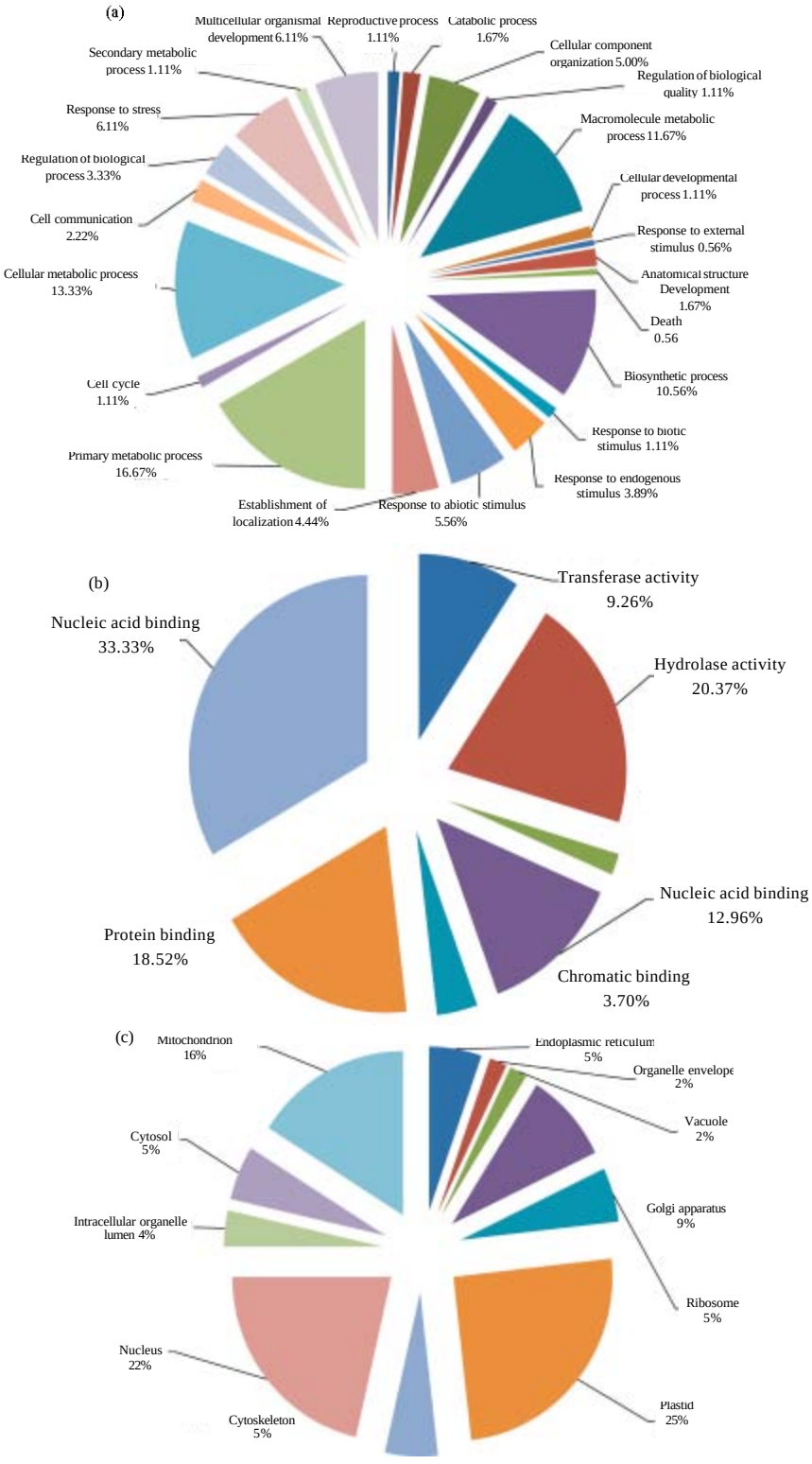


Fig. 3(a-c): Pie chart representations of GO-annotation results for (a) Biological process, (b) Molecular function and (c) Cellular component categories

Table 3: Genetic diversity parameters of kelampayan by EST-SSR analysis

| Mother tree | N | P (% P) | $A_a$ | $A_e$ | Mean heterozygosity $H_o$ | Mean heterozygosity $H_e$ |
|---|---|---|---|---|---|---|
| T1 | 26 | 15 (83.3) | 4.16 | 2.75 | 0.222±0.309 | 0.516±0.291 |
| T2 | 26 | 17 (94.4) | 4.11 | 3.24 | 0.231±0.322 | 0.595±0.245 |

Mean sample per locus (N), No. of polymorphic loci (P), Percentage of polymorphic loci (%P), Average No. of alleles per locus ($A_a$), Effective No. of alleles per locus ($A_e$), Observed heterozygosity ($H_0$), Expected heterozygosity ($H_e$)

(Table 3). The average PIC value of the full set of kelampayan EST-SSRs was calculated to be 0.465 and 0.537 for mother trees T1 and T2, respectively. After excluding all the non-polymorphic loci, the average PIC value was 0.558 and 0.569 for mother tree T1 and T2, respectively. The highest PIC value was 0.876 (NCS05) which was higher compared to rubber species (Feng *et al*., 2009) and Persian wheat (Zhuang *et al*., 2011) with the highest PIC value 0.684 and 0.714, respectively. The average expected heterozygosity ($H_e$) was 0.516±0.291 and 0.595±0.245 and the average observed heterozygosity ($H_o$) was 0.222±0.309 and 0.231±0.322 for mother trees T1 and T2, respectively. The range for $H_e$ and $H_o$ were slightly lower than cacao (Lima *et al*., 2010) which varied from 0 to 0.70 and 0.05-0.67, respectively. All loci were significantly deviated from HWE ($p<0.05$). This might be due to the nature of the samples used in this study, which are half-sib family samples and therefore, this may lead to the presence of excess homozygotes. Apart from that, the presence of null alleles would influence the heterozygosity values as well (Ellis and Burke, 2007). Thus, Micro-Checker Version 2.2.3 (Van Oosterhout *et al*., 2004) was used to validate the data. The result showed that homozygote excess and 13 loci showed significant evidence of null alleles. A total of 8 and 12 loci from mother trees T1 and T2, respectively showed high frequency of null alleles (r = 0.2). Hence, the lower value of heterozygosity and deviation from HWE are highly connected to the excess of homozygosity and the presence of null alleles of the half-sib samples used in this study.

**DNA sequencing of EST-SSR amplicon:** The randomly selected EST-SSR amplicons with various sizes were cloned and sequenced to further validate the SSR core motif of each locus. EST-SSR markers NCS03, NCS08, NCS09 and NCS13 were homozygous locus with only 1 allele per sample, and therefore two different individual samples were selected, cloned and sent for sequencing. Meanwhile two alleles from the same sample were sent for sequencing for EST-SSR marker NCS15 as it was a heterozygous locus. The sequence alignment between alleles derived from homozygous and heterozygous loci is shown in Fig. 4 and 5, respectively. Insertions and deletions (INDELs) were detected in the regions of SSR motifs and these results had further justified the declaration that the polymorphisms of SSRs are depend on the diversification of the number of repeat motifs (Liewlaksaneeyanawin *et al*., 2004). Base substitutions were detected in the primer flanking region in NCS08 and NCS15, while INDELs in the flanking region were observed in NCS08. This phenomenon was also found in rubber tree (Feng *et al*., 2009) where a complex mutational pattern, which involved changes in the number of SSR repeat units, base substitutions and INDELs within flanking regions were detected. The INDELs in the region of the SSR motifs could cause frame shift mutation of the gene and this has led to the expression of fully different or truncated proteins (Yan *et al*., 2012).

**Parentage assignment analysis:** A total of 54.8 and 40.2% of alleles detected were probably originated from the mother trees T1 and T2, respectively based on the parentage analysis by using half-sib family samples (Table 4). The low percentage of allele contributions could be due to the presence of null alleles as 18.7 and 23.5% of null allele frequencies were detected in mother trees T1 and T2, respectively in the present study. Null allele causes mismatches between the parent-offspring pairs and this is one of the potential constraint in parentage assignment using SSR loci (Castro *et al*., 2004). In general, gene flow does occur among the selected and non-selected kelampayan mother trees as about half of the alleles were potentially originated from other parents. This result was in agreement with the results obtained from the NJ tree constructed upon allele sharing between individuals (Fig. 6). There is a strong tendency that individuals with high similarity of genetic information will be clustered together in a group (Robichaud *et al*., 2006). Only a total of 32 and 28% of the kelampayan progenies were clustered together with its mother trees T1 and T2, respectively meanwhile others were grouped together in other three clusters as shown in Fig. 6. This indicates that there is a high probability for kelampayan to be predominantly outcrossed. This is supported by the high PIC values obtained in the present study. As reported by Glemin *et al*. (2006), the polymorphism rate would influence the outcrossing mating system rate for parentage analysis evaluation as the polymorphism rate for self-pollinated species is generally low within population.

**Transferability analysis:** The 18 newly designed kelampayan EST-SSR markers were tested against *Ixora caesia*, *Gardenia jasminoides*, *Mussaenda erythrophylla*, *Morinda citrifolia* and *Coffea canephora*. These are cross-genera species for kelampayan. PCR products were successfully obtained in

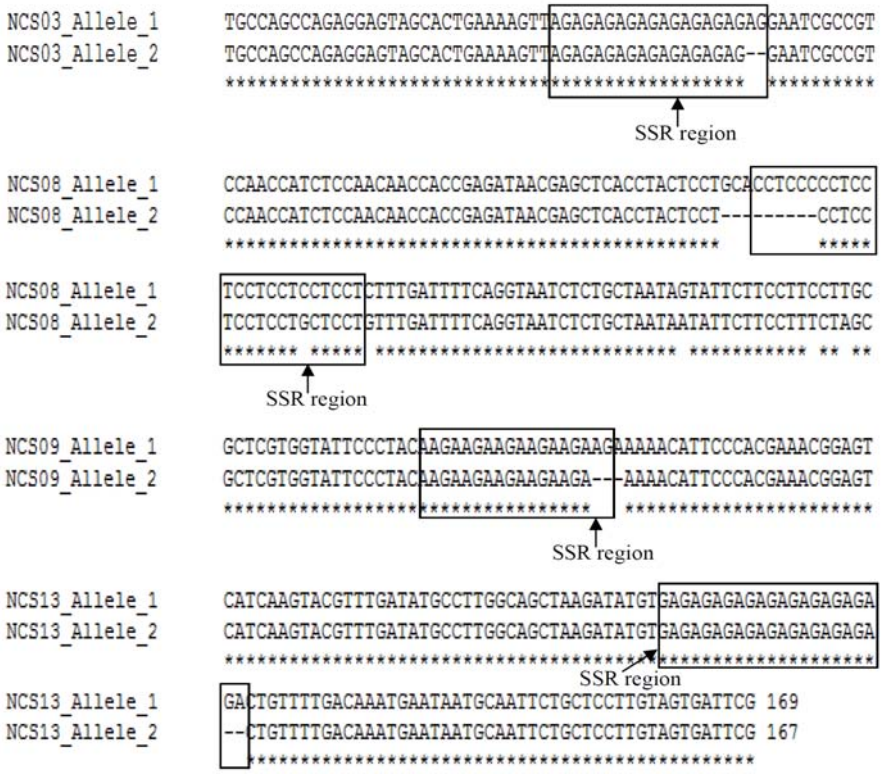


Fig. 4: DNA sequence alignment of EST-SSR alleles from a homozygous locus

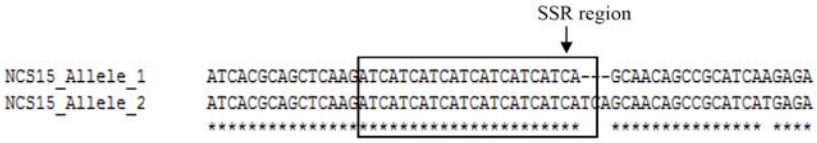


Fig. 5: DNA sequence alignment of EST-SSR alleles from a heterozygous locus

Table 4: Parentage assignment analysis of kelampayan by EST-SSR analysis

| | Mother tree T1 | | | Mother tree T2 | | |
|---|---|---|---|---|---|---|
| Locus | MT | Seedlings* | PA (%) | MT | Seedlings* | PA (%) |
| NCS01 | A, B | **A, B,** C, D, E | 48.9 | B | A, **B,** C, D, E | 4.9 |
| NCS02 | A | **A,** B | 96.0 | A | **A,** B | 68.0 |
| NCS03 | D | A, B, C, **D,** E | 24.0 | C | A, B, **C,** D | 32.0 |
| NCS04 | B | A, **B,** C | 36.0 | B | A, **B** | 72.0 |
| NCS05 | A, F | **A,** B, C, D, H, I, J, K, L | 20.0 | D, M | A, B, C, **D,** E, F, G, K, L, **M** | 24.0 |
| NCS06 | D, G | A, B, C, **D,** E, F, **G,** H | 46.2 | B | A, **B,** C, D, E, G, H | 17.6 |
| NCS08 | C | A, B, **C,** D | 52.0 | D | A, B, C, **D** | 4.0 |
| NCS09 | A | **A** | 100.0 | A | **A,** B | 91.3 |
| NCS10 | A, C | **A,** B, **C,** D, E | 48.6 | B, D | A, **B,** C, **D,** E | 34.1 |
| NCS12 | B | A, **B,** C, D, F, G | 35.1 | A | **A,** B, C, E, F, G | 20.5 |
| NCS13 | A | **A,** B, C | 60.0 | B | **B,** C | 56.0 |
| NCS14 | C | D, E, F | 0.0 | E | A, B, C, D, **E** | 16.0 |
| NCS15 | B | A, **B,** C, D | 43.3 | A | **A,** B | 24.0 |
| NCS16 | A | **A** | 100.0 | A | A | 100.0 |
| NCS17 | B | **B** | 100.0 | B | A, **B** | 64.0 |
| NCS18 | C, F | B, **C,** D, **F,** G | 69.0 | C, F | A, B, C, D, E, **F** | 37.0 |
| NCS20 | A | **A,** B, C | 34.5 | B | A, **B,** C | 26.3 |
| NCS22 | B, E | A, **B,** C, D, **E** | 72.4 | A, E | **A,** B, C, D, **E,** F | 32.3 |
| Mean | | | 54.8 | | | 40.2 |

MT: Mother tree, PA: Parentage allele contribution in percentage (%), *Bold alleles are probably originated from mother trees

Table 5: Cross-genera species amplification by using kelampayan EST-SSR markers

| EST-SSR loci | *Ixora caesia* | *Gardenia jasminoides* | *Mussaenda erythrophylla* | *Morinda citrifolia* | *Coffea canephora* | No. of species cross-amplifiable |
|---|---|---|---|---|---|---|
| NCS01 | + | + | + | + | + | 5 |
| NCS02 | + | + | - | + | + | 4 |
| NCS03 | + | + | - | + | - | 3 |
| NCS04 | + | + | - | + | + | 4 |
| NCS05 | + | + | + | + | + | 5 |
| NCS06 | + | + | - | - | + | 3 |
| NCS08 | + | + | - | + | + | 4 |
| NCS09 | - | + | - | + | + | 3 |
| NCS10 | + | + | - | + | + | 4 |
| NCS12 | + | - | - | + | + | 3 |
| NCS13 | - | + | - | + | + | 3 |
| NCS14 | - | + | - | - | - | 1 |
| NCS15 | + | + | - | - | + | 3 |
| NCS16 | + | + | - | + | + | 4 |
| NCS17 | + | + | - | + | + | 4 |
| NCS18 | + | + | - | + | + | 4 |
| NCS20 | + | + | + | + | + | 5 |
| NCS22 | - | + | - | + | - | 2 |
| No. of positive amplification | 14 (77.78%) | 17 (94.44%) | 3 (16.67%) | 15 (83.33%) | 15 (83.33%) | |

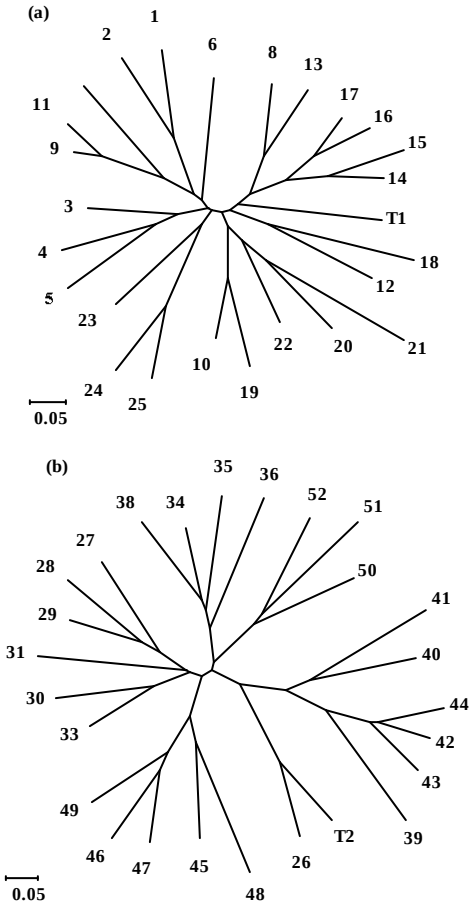


Fig. 6(a-b): (a) NJ tree of kelampayan mother tree 1 (T1) and (b) NJ tree of kelampayan mother tree 2 (T2) as displayed by MEGA version 3.1 (Kumar *et al*., 2004). No. 1 to 25 are progenies of T1 and No. 26 to 50 are progenies of T2

the cross transferability test, most of the loci were found to be transferable to *Gardenia jasminoides* (94.4%) followed by *Morinda citrifolia* (83.3%), *Coffea canephora* (83.3%) and *Ixora caesia* (77.8%) whereas least transferable to *Mussaenda erythrophylla* (16.7%) (Table 5). A total of three out of 18 (NCS01, NCS5 and NCS20) of the loci were transferable to all the five species across genera and only two loci (NCS14 and NCS22) showed amplification in less than two species. Hence, the gene-derived EST-SSR sequences from kelampayan were conserved between the five species from different genera. The transferability reduces as the genetic distance between the species increases (Liewlaksaneeyanawin *et al*., 2004). Hence, kelampayan has high possibility to be genetically closer to *G. jasminoides* and genetically distantly related to *M. erythrophylla*.

## CONCLUSION

In conclusion, the present study revealed a rapid and simple way to obtain EST-SSR markers from an EST database. To the best of our knowledge, this is the first report on the development of EST-SSR markers in kelampayan and we hope these markers could pave the way for exploiting the genotype data for comparative genome mapping, association genetics, population genetics studies and molecular breeding of kelampayan and other indigenous tropical tree species in future.

## ACKNOWLEDGMENTS

The authors would like to thank all the lab assistants and foresters involved in this research project for their excellent field assistance in sample collection. This study

is part of the joint Industry-University Partnership Programme, a research programme funded by the Sarawak Forestry Corporation (SFC) and Universiti Malaysia Sarawak (Grant No. 02(DPI09)832/2012(1), RACE/a(2)/884/2012(02) and GL(F07)/06/2013/STA-UNIMAS(06)).

## REFERENCES

Berube, Y., J. Zhuang, D. Rungis, S. Ralph, J. Bohlmann and K. Ritland, 2007. Characterization of EST-SSRs in loblolly pine and spruce. Tree Genet. Genomes, 3: 251-259.

Blair, M.W., N. Hurtado, C.M. Chavarro, M.C. Munoz-Torres and M.C. Giraldo *et al*., 2011. Gene-based SSR markers for common bean (*Phaseolus vulgaris* L.) derived from root and leaf tissue ESTs: An integration of the BMc series. BMC Plant Biol., Vol. 11 10.1186/1471-2229-11-50

Castro, J., C. Bouza, P. Presa, A. Pino-Querido and A. Riaza *et al*., 2004. Potential sources of error in parentage assessment of turbot (*Scophthalmus maximus*) using microsatellite loci. Aquaculture, 242: 119-135.

Ellis, J.R. and J.M. Burke, 2007. EST-SSRs as a resource for population genetic analyses. Heredity, 99: 125-132.

Feng, S.P., W.G. Li, H.S. Huang, J.Y. Wang and Y.T. Wu, 2009. Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). Mol. Breed., 23: 85-97.

Gao, L., J. Tang, H. Li and J. Jia, 2003. Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol. Breed., 12: 245-261.

Glemin, S., E. Bazin and D. Charlesworth, 2006. Impact of mating systems on patterns of sequence polymorphism in flowering plants. Proc. R. Soc. B, 273: 3011-3019.

He, G., R. Meng, M. Newman, G. Gao, R.N. Pittman and C.S. Prakash, 2003. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). BMC Plant Biol., Vol. 3 10.1186/1471-2229-3-3

Ho, W.S., R. Wickneswari, M.C. Mahani and M.N. Shukor, 2006. Comparative genetic diversity (polymorphisms) studies of *Shorea curtisii* Dyer ex King (Dipterocarpaceae) using SSR and DAMD markers. J. Trop. For. Sci., 18: 557-565.

Ho, W.S., S.L. Pang, P. Lau and I. Jusoh, 2011. Sequence variation in the *cellulose synthase* (*SpCesA1*) gene from *Shorea parvifolia* ssp. *parvifolia* mother trees. Pertanika J. Trop. Agric. Sci., 34: 317-323.

Ho, W.S., S.L. Pang, P.S. Lai, S.Y. Tiong and S.L. Phui *et al*., 2010. Genomics studies on plantation tree species in Sarawak. Proceedings of the International Symposium on Forestry and Forest Products 2010: Addressing the Global Concerns and Changing Societal Needs, October 5-7, 2010, Kuala Lumpur, Malaysia, pp: 172-182.

Joker, D., 2000. *Neolamarckia cadamba* (Roxb.) Bosser (*Anthocephalus chinensis* (Lam.) A. Rich. ex Walp.). Seed Leaflet No. 17, September 2000, Danida Forest Seed Centre, Denmark, pp: 1-2.

Kantety, R.V., M. La Rota, D.E. Matthews and M.E. Sorrells, 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol. Biol., 48: 501-510.

Kumar, S., K. Tamura and M. Nei, 2004. $MEGA_3$: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform., 5: 150-163.

Lau, E.T., W.S. Ho and A. Julaihi, 2009. Molecular cloning of *cellulose synthase* gene, *SpCesA1* from developing xylem of *Shorea parvifolia* spp. *parvifolia*. Biotechnology, 8: 416-424.

Li, Y.C., A.B. Korol, T. Fahima and E. Nevo, 2004. Microsatellites within genes: Structure, function and evolution. Mol. Biol. Evol., 21: 991-1007.

Liang, X., X. Chen, Y. Hong, H. Liu, G. Zhou, S. Li and B. Guo, 2009. Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species. BMC Plant Biol., Vol. 9 10.1186/1471-2229-9-35.

Liewlaksaneeyanawin, C., C.E. Ritland, Y.A. El-Kassaby and K. Ritland, 2004. Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor. Applied Genet., 109: 361-369.

Lima, L.S., K.P. Gramacho, J.L. Pires, D. Clement and U.V. Lopes *et al*., 2010. Development, characterization, validation and mapping of SSRs derived from *Theobroma cacao* L., *Moniliophthora perniciosa* interaction ESTs. Tree Genet. Genomes, 6: 663-676.

Liu, K. and S.V. Muse, 2005. Power Marker: An integrated analysis environment for genetic marker analysis. Bioinformatics, 21: 2128-2129.

Luro, F.L., G. Costantino, J. Terol, X. Argout and T. Allario *et al*., 2008. Transferability of the EST-SSRs developed on Nules clementine (*Citrus clementina* Hort ex Tan) to other *Citrus* species and their effectiveness for genetic mapping. BMC Genomics, Vol. 9 10.1186/1471-2164-9-287.

Marconi, T.G., E.A. Costa, H.R. Miranda, M.C. Mancini and C.B. Cardoso-Silva *et al*., 2011. Functional markers for gene mapping and genetic diversity studies in sugarcane. BMC Res. Notes, Vol. 4 10.1186/1756-0500-4-264

Pashley, C.H., J.R. Ellis, D.E. Mccauley and J.M. Burke, 2006. EST databases as a source for molecular markers: Lessons from *Helianthus*. J. Heredity, 97: 381-388.

Poncet, V., M. Rondeau, C. Tranchant, A. Cayrel, S. Hamon, A. de Kochko and P. Hamon, 2006. SSR mining in coffee tree EST databases: Potential use of EST-SSRs as markers for the *Coffea* genus. Mol. Genet. Genomics, 276: 436-449.

Robichaud, R.L., J.C. Glaubitz, O.E. Jr. Rhodes and K. Woeste, 2006. A robust set of black walnut microsatellites for parentage and clonal identification. New For., 32: 179-196.

Soerianegara, I. and R.H.M.J. Lemmens, 1993. Plant Resources of South-East Asia No 5(1). Timber Trees: Major Commercial Timbers. Pudoc Scientific Publisher, Wangeningen, Netherlands, pp: 107.

Tchin, B.L., W.S. Ho, S.L. Pang and J. Ismail, 2011. Gene-associated single nucleotide polymorphism (SNP) in *Cinnamate 4-Hydroxylase* (C4H) and *Cinnamyl alcohol dehydrogenase* (CAD) genes from *Acacia mangium* superbulk trees. Biotechnology, 10: 303-315.

Tchin, B.L., W.S. Ho, S.L. Pang and J. Ismail, 2012. Association genetics of the *Cinnamyl alcohol dehydrogenase* (*CAD*) and *Cinnamate 4-hydrolase* (*G4H*) genes with basic wood density in *Neolamarckia cadamba*. Biotechnology, 11: 307-317.

Van Oosterhout, C., W.F. Hutchinson, D.P.M. Wills and P. Shipley, 2004. Micro-Checker: Software for identifying and correcting genotyping errors in microsatellite data. Mol. Ecol. Notes, 4: 535-538.

Varshney, R.K., A. Graner and M.E. Sorrells, 2005. Genic microsatellite markers in plants: Features and applications. Trends Biotechnol., 23: 48-55.

Xu, M., Y. Sun and H. Li, 2010. EST-SSRs development and paternity analysis for *Liriodendron* spp. New For., 40: 361-382.

Yan, M., X. Dai, S. Li and T. Yin, 2012. A meta-analysis of EST-SSR sequences in the genomes of pine, poplar and eucalyptus. Tree Genet. Mol. Breed., 2: 1-7.

Yasodha, R., R. Sumathi, P. VChezhian, S. Kavitha and M. Ghosh, 2008. *Eucalyptus* microsatellites mined *in silico*: Survey and evaluation. J. Genet., 87: 21-25.

Yeh, F.C., R.C. Yang and T. Boyle, 1997. POPGENE Version 1.32: Microsoft Window-Based Freeware for Population Genetic Analysis. University of Alberta, Edmonton, Canada.

Yu, J.K., T.M. Dake, S. Singh, D. Benscher, W.L. Li, B. Gill and M.E. Sorrells, 2004. Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. Genome, 47: 805-818.

Zhuang, P.P., Q.C. Ren, W. Li and G.Y. Chen, 2011. Genetic diversity of Persian wheat (*Triticum turgidum* sp. carthlicum) accessions by EST-SSR markers. Am. J. Biochem. Mol. Biol., 1: 223-230.